基于动态 BLSTM 和 CTC 的濒危语言语音识别研究 *

于重重 [1]，陈运兵 [1]，孙沁瑶 [1]，刘　畅 [1]，徐世璇 [2]，尹蔚彬 [2]

(1. 北京工商大学 计算机与信息工程学院，北京 100048; 2. 中国社会科学院民族学与人类学研究所，北京 100081)

摘　要：针对低资源的濒危语言进行了端到端语音识别模型的建立与研究，能够为濒危语言的保护和传承探索出新的途径。采用动态双向长短时记忆网络与连接时序分类模型构造端到端的语音识别系统，在做音素级别的识别训练时，传入模型的数据批量大小根据训练模型做自适应调整，不仅能够加快收敛速度，而且能够提高模型的泛化性。通过修改网络层次与结构参数，并提取不同的语音特征进行模型对比，实验结果表明在两种濒危语言——吕苏语和土家语的数据集上均取得了较好的识别效果。

关键词：濒危语言语音识别；端到端；动态双向长短时记忆网络；连接时序分类模型

中图分类号：TP391.42　　doi: 10.3969/j.issn.1001-3695.2018.05.0291

Endangered languages speech recognition based on dynamic BLSTM and CTC

Yu Chongchong[1], Chen Yunbing[1], Sun Qinyao[1], Liu Chang[1], Xu Shixuan[2], Yin Weibin[2]

(1. *College of Computer & Information Engineering*, *Beijing Technology & Business University*, *Beijing* 100048, *China*; 2. *Institute of Ethnology & Anthropology*, *Chinese Academy of Social Sciences*, *Beijing* 100081, *China*)

Abstract: In view of the low resource of endangered languages, the establishment and research of end-to-end speech recognition model can explore new ways for the protection and transmission of endangered languages. this paper combined dynamic bi-directional long short-term memory network and connectionist temporal classification model into an end-to-end speech recognition model. When performing phoneme-level recognition training, the batch size of the data passed into the model can be adaptively adjusted according to the training model, which not only speeds up the convergence but also improves the generalization of the model. By adjusting the hierarchy of the deep neural network and extracting different phonetic features for model comparison, the experimental results show that both the endangered languages - Lvsu and Tujia have good recognition results.

Key words: endangered languages speech recognition; end to end; dynamic bi-directional long short-term memory; connectionist temporal classification

0　引言

语音信号是一种非平稳时序信号，其形成和感知的过程就是一个复杂信号的处理过程，而语音识别可视为一个序列到序列的分类问题[1]，即声学观测序列 $X=(x_1,x_2,\cdots,x_T)$ 被映射到字符序列 $W=(w_1,w_2,\cdots,w_N)$ 上,其中 T 为时间， N 为字符个数，解决概率 $P(W\mid X)$ 的问题。在时序分类任务中，常用的方法是输入数据与给定标签必须要在时间上达到帧级别的对齐，只有这样才能使用隐马尔可夫模型[2]（hidden Markov model, HMM）按帧进行训练。然而，逐帧训练输出的是单帧概率，对于时序问题来说，输出序列的概率远比输出单帧的概率重要得多。因此，针对此问题，在基于深度学习的语音识别领域中，端到端的语音识别技术[3~5]已成为国内外近期研究的热点之一。

文献[6~8]提出由长短时记忆（long short-term memory，LSTM）网络和连接时序分类（connectionist temporal classification，CTC）结合而成的端到端的语音识别系统模型，该模型直接对一段语音的音素序列或者绑定的音素（context-dependent phone,CD-phone）序列与对应的语音特征序列进行序列层面建模，不需要利用 HMM 进行强制对齐得到帧级别的标注，可以取得相比于传统 LSTM-HMM 声学模型更好的性能。

收稿日期：2018-05-02；**修回日期：**2018-06-20　**基金项目：**国家教育部人文社会科学研究规划基金资助项目（16YJAZH072）；国家社会科学基金重大项目（14ZDB156）

作者简介：于重重（1971-），女，辽宁丹东人，教授，博士，主要研究方向为智能信息处理、模式识别与机器学习（chongzhy@vip.sina.com）；陈运兵（1992-），男，山东菏泽人，硕士研究生，主要研究方向为机器学习、智能信息处理；孙沁瑶（1992-），男，吉林人，硕士研究生，主要研究方向为机器学习、智能信息处理；刘畅（1998-），男，四川遂宁人，本科，主要研究方向为模式识别与机器学习；徐世璇（1954-），女，浙江宁波人，研究员，博导，主要研究方向为少数民族语言；尹蔚彬（1969-），女，河北沧州人，副研究员，博士，主要研究方向为藏缅语族语言研究、藏区语言生态.

文献[9]基于注意机制（attention）的端到端模型，直接实现从语音声学特征序列到最终句子级的音素序列、字符序列或词序列的输出,但是在大词汇量连续语音识别任务上，该方法的性能目前和最好的语音识别系统的性能还有一定的差距。

目前国内外自动语音识别技术多数是依赖于大量的数据资源，而濒危语言是指使用人数越来越少的、行将灭绝的语言，可采集的语音数据量非常有限，属于低资源语音识别。濒危语言多数没有文字，以口语的形式存在，母语人的数量少，导致数据不易收集,因此对濒危语言的自动语音识别有很大挑战性。据统计，我国少数民族使用的语言在 130 种以上，近一半处于衰退状态，当前我国有几十种语言处于濒危状态，这种趋势仍在持续,甚至有所加剧[10],对濒危语言的识别与保护有助于维护文化的多样性。

针对濒危语言语音识别的研究，文献[11]提出结合 CTC 技术和藏语语言学知识，使用绑定的三音子(tri-phone)作为建模单元，解决建模单元的稀疏性问题，但训练语料的稀疏性严重降低了声学模型的区分度鲁棒性。文献[12]将瓶颈特征及其与 MFCC 的复合特征用于藏语拉萨语连续语音识别任务中，代替传统的 MFCC 特征进行 GMM-HMM 声学建模，虽然识别准确率得到了一定的提升，但是使用的仍是传统语音识别方法。文献[13]针对低资源条件下带标注训练数据较少的问题，提出基于 i-vector 特征的 LSTM 递归神经网络系统，并在 OPEN KWS 2013 标准数据集上字节错误率获得了显著的下降，但是缺少对 LSTM 网络进行优化。文献[14]在使用 CTC 网络时加入 Attention 模型，有效地完成了低资源语言的关键词搜索和语音识别，但是语音识别效果较差。由于端到端的语音识别系统不仅在训练过程中自动学习声学特征和标注序列的对应关系，不需要强制状态对齐等一系列繁琐的步骤，而且减少了对发音词典的要求。

本文实验数据为两种濒危语言—吕苏语和土家语，针对濒危语言的低资源性，对端到端的语音识别模型进行研究。在模型的编码阶段，采用动态双向长短时记忆（dynamic bi-directional LSTM, DBLSTM）网络对长序列建模，DBLSTM 网络是由两个单向 LSTM 上下叠加在一起组成，其输出由这两个 LSTM 网络的状态共同决定，而在每一次训练时传入模型的批量大小是可变的，并根据训练模型进行自适应调整，这样不仅能够有效地挖掘语音信号的帧间先验信息，而且可以提高模型的泛化性。而在解码阶段，采用 CTC 自动学习并优化输入输出序列的对应关系，得到整体序列的概率，从而减少了标签预划定的冗杂工作。

1　双向长短期记忆网络

在时序模型中，循环神经网络[15]（recurrent neural network,RNN）,其应用场景十分广泛，可用于语音识别、机器翻译、看图说话、问答系统等领域。循环神经网络自身的结构特点已使得它能够较好地挖掘利用序列数据的信息，即具有记忆性，在时间序列数据学习方面具有强大的建模能力，能够以一种灵活的方式结合数据的背景信息，对即使发生局部畸变的数据也可以有效地完成学习任务。训练 RNN 的方法是在传统的反向传播（back propagation,BP）上加了时间的考量，称为 BPTT（back propagation through time）。实际中如果记忆的窗口太长，RNN 会存在训练不稳定，梯度消失或爆炸等问题。为了克服 RNN 的记忆缺陷，Graves 提出 LSTM 网络[16]，该网络结构采用了大量记忆单元(cell)和复杂的信息流处理手段，用于记忆上下文信息，从而对语音的长时相关性进行建模。

在 LSTM 网络中，每个神经元是一个“记忆细胞”，细胞里面有一个“输入门”、一个“遗忘门”和一个“输出门”，可以选择性记忆历史信息。输入门决定何时让输入进入细胞单元，遗忘门决定何时应该记住前一时刻的记忆，输出门决定何时让记忆流入到下一时刻。LSTM 在 t 时刻按照如下式子进行计算。

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \tag{2}$$

$$c_t = f_t c_{t-1} + i_t \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o) \tag{4}$$

$$h_t = o_t \phi(c_t) \tag{5}$$

$$y_t = W_{yh}h + b_y \tag{6}$$

其中：i_t、f_t、c_t、o_t、h_t 是分别是输入门、遗忘门、记忆单元、输出门和隐藏层状态，W 为各部分的权值矩阵，如 W_{ix} 为输入门与输入层之间的权值矩阵，b 为各部分的偏置矩阵，σ 是 sigmoid 函数，ϕ 为神经元激活函数，如 tanh 等。

语音信号不仅是一种非平稳的随机信号，而且是一种前后相关性较强的信号，如何有效地对长时序动态相关性进行建模至关重要。为了充分利用未来的上下文信息，Graves 等将双向长短期记忆网络（bi-directional LSTM, BLSTM）应用于语音识别，该网络结构由两个单向 LSTM 上下叠加在一起组成，其输出由这两个 LSTM 网络的状态共同决定，可以提供给输出层完整的过去和未来的上下文信息。

2　连接时序分类

连接时序分类 CTC 是由 Graves 等人于 2006 年提出来的一种时序分类算法[17]，CTC 与传统的方法不同，其不需要标签在时间上的帧级别对齐就可以进行训练，对输入数据的任一时刻做出的预测不是很关心，而其重点关注的是整体上输出是否与标签一致，CTC 输出的是整体序列的概率，从而减少了标签预划定的冗杂工作。CTC 网络输出层还包含一个空(blank)节点，这个 blank 标注主要是为了对静音、字间停顿、字间混淆进行建模。因此，CTC 很善于处理时序分类问题。

设给定输入序列 $X=(x_1,x_2,\cdots,x_T)$，时间为从 1 到 T 时刻，CTC 网络按公式(1)~(6)计算其对应的输出序列 $Y=(y_1,y_2,\cdots,y_T)$，其中 $y_i=(y_i^1,y_i^2,\cdots,y_i^K)$，$i=1,2,\cdots,K$ 为第 i 帧的条件概率分布，则 softmax 层的输出为

$$P(k \mid t,x)=\frac{\exp(y_t^k)}{\sum_{k=1}^{L}\exp(y_t^{k'})} \tag{7}$$

其中：K 为所有标签个数，即 CTC 网络输出层结点个数 K。

对于 T 帧声学输入，CTC 网络学习得到长度为 T 的标注序列 π 的概率为

$$P(\pi \mid x)=\prod_{t=1}^{T}P(\pi_t \mid t,x) \tag{8}$$

对于给定的标注序列 μ，由于 blank 插入的位置不同及非 blank 标注重复性的存在，π 与 μ 存在多对一的关系。因此可将目标函数重写如下：

$$P(\mu \mid x)=\sum_{\pi\in B^{-1}(y)}P(\pi \mid x) \tag{9}$$

其中：$\mu=B(\pi)$ 为映射函数，即给定参考标注 μ 目标函数定义如下：

$$CTC(x)=-\log P(\mu \mid x) \tag{10}$$

从上述 CTC 网络的训练过程不难看出，CTC 网络解码的最佳路径就是在给定输入序列的情况下，找到概率最大的输出序列：

$$\mu\approx B(\pi^*),\quad \pi^*=\arg\max_{\pi} P(\pi \mid x) \tag{11}$$

其中：π^* 为 T 帧输入序列的后验概率输出的最大值对应的标注序列。

3　濒危语言语音识别模型

3.1　端到端的 DBLSTM-CTC 模型

为了解决濒危语言低资源问题，实现更好的语音识别模型，本文采用 DBLSTM 网络与 CTC 模型结合共同构造端到端的语音识别系统。图 1 给出了基于 DBLSTM-CTC 声学模型的语音识别系统框图，本文中所使用的濒危语言数据均没有文字，模型输入的时序为每帧语音特征，输出的时序为国际音标。首先对实验音频数据进行一系列的预处理，然后分析语音的频谱并提取相关特征，接着采用 DBLSTM 网络对长序列进行建模，充分挖掘上下文信息。

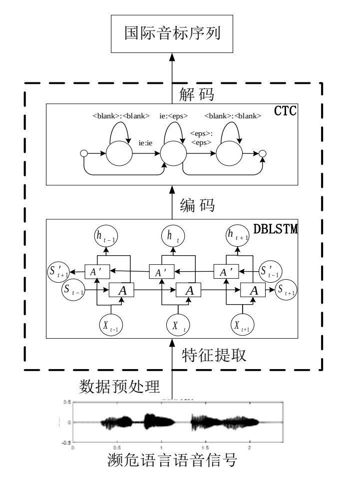


图 1　基于端到端的濒危语言语音识别模型

在解码阶段，由于 CTC 可以被视为一种能够直接优化输入序列与输出目标序列似然度的目标函数，在此目标函数下，CTC 在训练过程中自动学习并优化输入输出序列的对应关系。CTC 网络的输出层为 softmax 层，节点个数与标注序列的个数相同。而在解决标注存在叠字的问题上，blank 节点起到重要作用。在语音识别中的一帧数据很难给出一个 label，但是几十帧就容易判断出对应的发音 label。在 CTC 网络中，正是由于 blank 节点的存在，所以才可以采取跳帧的方法。CTC 的输出和 label 满足如下的等价关系：

$$F(i-ie-)=F(-ii--ie)=iie \tag{12}$$

其中：“i”和“e”为濒危语言国际音标，“-”为 blank，由式（12）可以看出，多个输出序列可以映射到一个输出。因此 CTC 不仅能够加快解码速度，而且在训练过程中自动优化输入输出序列的对应关系。

3.2　动态双向长短期记忆网络

由于濒危语言的低资源性导致语音数据存在的数据稀疏问题。在编码阶段，每一次训练时传入模型的批量大小是可变的，本文称这种双向 LSTM 模型为动态双向长短时记忆网络。DBLSTM 网络是由两个单向 LSTM 上下叠加在一起组成，其输出由这两个 LSTM 网络的状态共同决定，在做音素级别的识别训练时，batch 大小能够根据训练模型做自适应调整，首先根据硬件配置设定 batch 大小上限，给定初始 batch 大小（即下限），以 2 为基本增减单位，通过计算前一次训练的损失函数均值与方差来判断后一次训练 batch 大小的增大或减小，以适应模型权值变化的统计特性，这样不仅能够加快收敛速度，而且能够提高模型的泛化能力。图 2 展示了在训练时设置 batch 大小固定为 16 和给定 batch 初始值为 10 并对 batch 自适应调整时的损失函数值变化情况。

通过图 2 可以看出，当给定 batch 初始值并做自适应调整，相对于设置 batch 固定大小时，模型在训练时迭代相同次数时损失函数值更小且收敛速度更快。

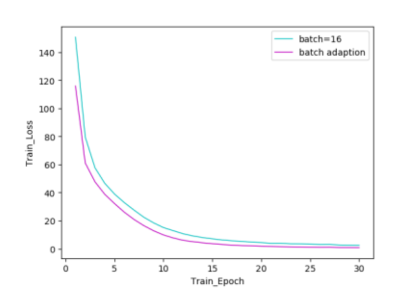


图 2　设定 batch=16 时与 batch 自适应时的损失函数值变化

本文选用适应性动量估计算法[18]（adaptive moment estimation, Adam）作为模型的优化算法，该算法结合了 AdaGrad 和 RMSProp 算法的优点，Adam 不仅可以基于一阶矩均值计算适应性参数学习率，它同时还充分利用了梯度的二阶矩均值，能够对不同参数计算适应性学习率并且占用较少资源。

4 实验及结果分析

4.1 实验数据

本文实验使用以下两种濒危语言的语料数据:吕苏语包括 15 篇口语短篇语料，共有 6257 个句子和 4 149 个词汇，总计时长为 2 小时 52 分 20 秒。土家语语料包括 3 篇口语短篇语料、300 核心词口语语料和 2 000 个主要词的口语语料、和部分语法例句口语语料，共有 2130 个句子和 10029 个词汇，总计时长为 5 小时 9 分 15 秒。

利用 ELAN 软件濒危语言口语语料对进行人工标注和存储。ELAN 主要工作流程包括三个部分：分割、转写、翻译。首先根据说话人的语音间隔将有内容的语音信息分离出来，其次根据语音和记音内容进行人工转写标注，最后把转写出对的词汇根据汉语语法信息人工串成一句完整通顺的句子。

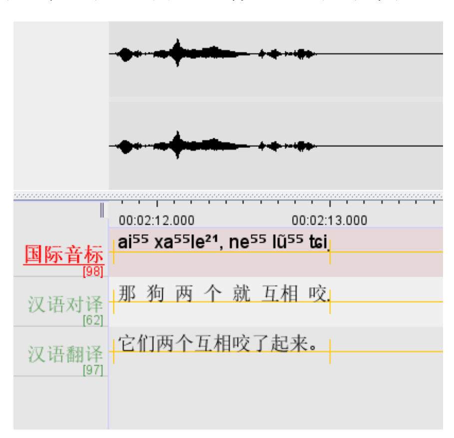


图 3　原始濒危语言语料标注内容

由图 3 可看出，使用 ELAN 软件标注的内容包括：口语-国际音标、口语-汉语对译、口语-汉语翻译。

吕苏语的音位系统由声母和韵母构成[19~21]，其中声母部分由 38 个单辅音和 27 个复辅音组成，韵母部分由 18 个单元音和 12 个复元音组成。同样地，土家语的音位系统也是由声母和韵母构成，21 个声母中包括两个半元音声母，韵母由 6 个单元音、复元音 11 个和 8 个鼻化元音组成。

4.2 实验平台

本文实验采用的服务器设备为 Dell PowerEdge R730，其中处理器为 Intel(R) Xeon(R) CPU E5-2643 v3 @3.40 GHz，实验环境为在 Ubuntu 16.04 系统上安装的深度学习框架 TensorFlow 1.1.0，Cuda 8.0。

4.3 实验模型参数选取

本文在训练过程中，初步设置 DBLSTM 模型参数如下：batch 大小初始为 10，隐层个数为 2，隐层节点个数为 256。图 4 和图 5 展示了相同参数下不同学习率（lr=0.005,0.001,0.0001）对训练语音识别系统的损失函数变化和模型精度的影响。

1）吕苏语语料实验部分

由图 4 和 5 可以看出，随着迭代次数的增加，当学习率 lr=0.001 时，最终获得的损失函数较小，同时训练时的错误识别率也较小,说明学习率在选取时不是越小越好，需要进行不同学习率模型结果的对比。

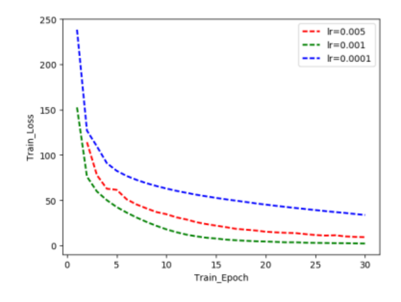


图 4　吕苏语在不同学习率下训练时的损失函数变化情况

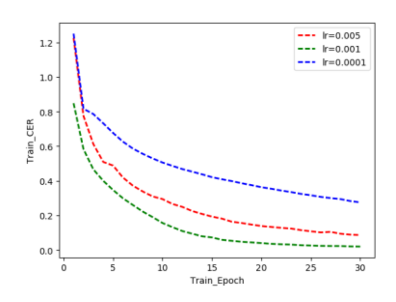


图 5　吕苏语在不同学习率下训练时的错误识别率

2) 土家语语料实验部分

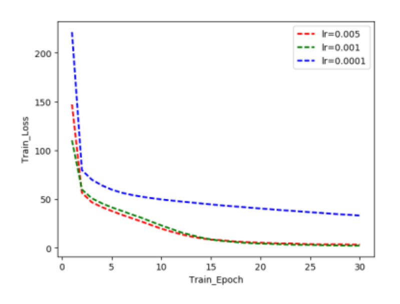


图 6　土家语在不同学习率下训练时的损失函数变化情况

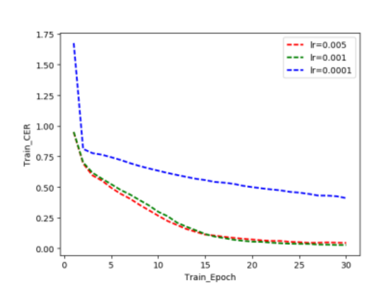


图 7　土家语在不同学习率下训练时的错误识别率

由图 6 和 7 也可以看出，随着迭代次数的增加，当学习率 lr=0.001 时，最终获得的损失函数较小，同时训练时的错误识别率也较小，因此本文实验中的学习率均取值为 0.001。

通过修改网络结构中的隐层个数、激活函数等参数，最后得到语音识别实验的最佳识别率，此时相应的 DBLSTM 模型参数分别是：隐层个数为 3，隐层节点个数为 512，学习率为 0.001。

4.4 实验分析

为了验证本文基于 DBLSTM 网络与 CTC 模型结合共同构造端到端的语音识别系统的优势，将 DBLSTM 模型替换为 BLSTM 模型进行对比，设置模型参数分别是：batch 大小初始为 10，隐层个数为 3，隐层节点个数为 512，学习率为 0.001，

即以相同参数分别对吕苏语和土家语语料进行不同语音识别模型的实验，并提取不同的语音特征-MFCC (Mel-frequency cepstral coefficients)和 FBank (filter bank)进行对比，实验结果如下：

1）吕苏语实验部分

表 1　吕苏语语料在不同系统上的实验结果

识别模型	语音特征	识别率
BLSTM+CTC	MFCC	72.11%
	FBank	72.79%
DBLSTM+CTC	MFCC	72.51%
	FBank	73.19%

通过表 1 可以看出，在提取不同语音特征下，DBLSTM-CTC 模型的语音识别准确率显然都比 BLSTM-CTC 模型好。

2）土家语实验部分

表 2　土家语语料在不同系统上的实验结果

识别模型	语音特征	识别率
BLSTM+CTC	MFCC	46.31%
	FBank	47.67%
DBLSTM+CTC	MFCC	47.76%
	FBank	49.98

同样地，而通过表 2 可以看出，在提取不同语音特征下，DBLSTM-CTC 模型的语音识别准确率明显比 BLSTM-CTC 模型好。特别地，由于土家语语料中环境噪声较大，导致语音质量较差，因此识别率相对较低，因此后续工作将研究如何去除环境噪声改善语音质量以提高识别率。

总之，由表和图中的实验结果数据可知，在端到端的语音识别系统的编码阶段，DBLSTM 模型无论是损失变化还是语音识别率都明显有优于 BLSTM 模型，证明了 DBLSTM 结构在序列预测和序列标注任务中具有较好的性能，能够有效地对语音的长时相关性进行建模，并且泛化能力较好。另外，由于本文所使用濒危语言语料资源受限，以及本文实验语料缺少大语种语料库的标准性，直接采用国际音标作为输出建模，并为了保证模型的稳定性，删除了一些特殊音符，最后建立的濒危语言语音识别系统在吕苏语和土家语数据集上取得了一定的成果。本文对吕苏语和土家语语音识别实验的结果表明，采用端到端的 DBLSTM-CTC 模型不仅取得了有效的识别结果，而且为濒危语言的保护和传承探索出新的途径。

5　结束语

本文针对濒危语言的特定场景，采用 DBLSTM 网络与 CTC 模型结合共同构造端到端的语音识别系统，在模型训练时给定 batch 初始值并做自适应调整，不仅能够加快收敛速度，而且提高了模型的泛化性，实验结果表明在吕苏语和土家语数据集上均取得了较好的识别效果。虽然本文建立的濒危语言语音识别系统在吕苏语和土家语数据集上取得了一定的成果，但是最后语音识别效果并没有达到预期效果，因此后续会通过拓展语料等方法提高系统模型的稳定性和识别准确率。

参考文献：

[1] Mohamed A R, Dahl G, Hinton G. Deep belief networks for phone recognition [EB/. OL]. (2010-07) . http://www. cs. toronto. edu/~gdahl/papers/dbnPhoneRec. pdf.

[2] Rosenberg A, Audhkhasi K, Sethy A, *et al.* End-to-end speech recognition and keyword search on low-resource languages [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017: 5280-5284.

[3] Lu Liang, Kong Lingpeng, Dyer C, *et al.* Multi-task learning with CTC and segmental CRF for speech recognition [C]// Proc of INTERSPEECH. 2017: 954-958.

[4] Ochiai T, Watanabe S, Hori T, *et al.* Multichannel end-to-end speech recognition [J/OL]. (2017-03-14) . https://arxiv. org/pdf/1703. 04783. pdf

[5] Zhang Yu, Chan W, Jaitly N. Very deep convolutional networks for end-to-end speech recognition [J]. (2016-10-10) . https://arxiv. org/abs/1610. 03022.

[6] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks [C]// Proc of International Conference on Machine Learning. 2014: 1764-1772.

[7] Haşim Sak, Senior A, Rao K, *et al.* Learning acoustic frame labeling for speech recognition with recurrent neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2015: 4280-4284.

[8] Graves A, Jaitly N, Mohamed A R. Hybrid speech recognition with Deep Bidirectional LSTM [C]// Automatic Speech Recognition and Understanding. IEEE, 2014: 273-278.

[9] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J/OL]. (2016-05-19) . https://arxiv. org/pdf/1409. 0473. pdf

[10] 徐世璇. 我国濒危语言研究的历程和前景 [J]. 西北民族大学学报: 哲学社会科学版, 2015 (1): 83-90. (Xu Shixuan. The course and prospect of endangered language studies in China [J]. Journal of Northwest University for Nationalities: Philosophy and Social Science, 2015 (1): 83-90.)

[11] 王庆楠, 郭武, 解传栋. 基于端到端技术的藏语语音识别 [J]. 模式识别与人工智能, 2017, 30 (4): 359-364. (Wang Qingnan, Guo Wu, Xie Chuandong. Towards end to end speech recognition system for Tibetan [J]. Pattern Recognition and Artificial Intelligence, 2017, 30 (4): 359-364.)

[12] 周楠, 赵悦, 李要嫱, 等. 基于瓶颈特征的藏语拉萨话连续语音识别研究 [J]. 北京大学学报: 自然科学版, 2018, 54 (2): 249-254. (Zhou Nan, Zhao Yue, Li Yaoqiang, *et al.* Study on continuous speech recognition based on bottleneck features for Lhasa-Tibetan dialect [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2018, 54 (2): 249-254.)

[13] 黄光许, 田垚, 康健, 等. 低资源条件下基于 I-vector 特征的 LSTM 递归神经网络语音识别系统 [J]. 计算机应用研究, 2017, 34 (2): 392-396. (Huang Guangxu, Tian Yao, Kang Jian, *et al.* Long short term memory

recurrent neural network acoustic models using i-vector for low resource speech recognition [J]. Application Research of Computers, 2017, 34 (2): 392-396)

[14] Rosenberg A, Audhkhasi K, Sethy A, *et al.* End-to-end speech recognition and keyword search on low-resource languages [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017: 5280-5284.

[15] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks [J]. 2013, 38 (2003): 6645-6649.

[16] Graves A. Long Short-Term Memory [M]// Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012: 1735-1780.

[17] Graves A, Gomez F. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]// Proc of International Conference on Machine Learning. New York: ACM Press, 2006: 369-376.

[18] Kingma D P, Ba J. Adam: a method for stochastic optimization [J/OL]. (2017-01-30) . https://arxiv. org/pdf/1412. 6980. pdf

[19] 于重重，操镭，尹蔚彬，等. 吕苏语口语标注语料的自动分词方法研究 [J]. 计算机应用研究, 2017, 34 (5): 1325-1328. (Yu Chongchong, Cao Lei, Yin Weibin, Zhang Zeyu, *et al.* Automatic word segmentation on Lizu spoken annotation corpus [J] , Application Research of Computers, 2017, 34 (5): 1325-1328.)

[20] 林幼菁，尹蔚彬，王志. 吕苏语的助动词 [J]. 民族语文, 2014 (2) . (Lin Youjing, Yin Weibin, Wang Zhi. Helper verbs in Lvsu [J] , Minority Languages of China, 2014 (2)

[21] 徐世璇. 土家语空间概念的语法和语义表征 [J]. 民族语文, 2013 (1): 35-45. (Xu Shixuan. Grammatical and semantic representation of spatial concepts in the Tujia language [J]. Minority Languages of China, 2013 (1): 35-45) .